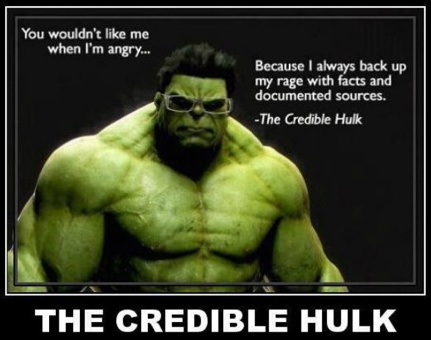
**AP Statistics – Summer Assignment**

This assignment is due at the end of the first week of class. It is considered your first test grade, which as a category is worth approximately 60% of your total grade. The assignment reinforces concepts from prior curricula and introduces a lot of the things we will be working on during the year. **The final page of the assignment is the ANSWER PAGE.** **This is the only page that needs to be turned in. All work must be shown and graphs need to be drawn accurately on the answer sheet.** Each topic contains data and instructions for what you need to do. You will need to use the website <https://stapplet.com> when directed. Any questions can be sent to [jhoiler@ccboe.com](mailto:jhoiler@ccboe.com).

Meet our mascot for the year… The Credible Hulk ☺

**Step 1** – Go to Youtube and watch this video: <https://youtu.be/sxQaBpKfDRk>

I hope this gives you a quick intro into the types of questions we are going to explore this year in AP Stat. Statistics is my favorite math course to teach, and I hope you learn enough to start asking important questions of your own.

Below, the first time I mention an important term, I will try to bold it and provide an explanation.

**Statistics** – the science and art of collecting, analyzing, and drawing conclusions from data.

***Topic 1 – Displaying Categorical Data***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Core** | |  |
| **Elective** |  | Math | English | TOTAL |
| Art | 2 | 4 |  |
| Music | 5 | 4 |  |
| P.E. | 4 | 3 |  |
| Language | 1 | 3 |  |
| Tech. | 4 | 0 |  |
|  | TOTAL |  |  |  |

I did a poll of my class last fall, and I asked them what their favorite core class, and what their favorite elective class was. I show the results in the two-way table on the right. Each box lists the number of people who chose that combination

1. On the answer sheet, fill in the total amount of students who chose each elective. (Add horizontally)

This data is called **CATEGORICAL** data because it sorts individuals into categories that don’t have mathematical meaning. Only the quantity of individuals in each category can be counted. The students are the **INDIVIDUALS** in the survey, and the two **VARIABLES** of interest are Favorite Core and Favorite Elective class.

1. Go to the website <https://stapplet.com> and choose the “One Categorical Variable, Single Group” option. Enter the categories of Favorite Elective and the total number of students who chose each. Create a bar chart and a pie chart and sketch them on the answer sheet. REMEMBER TO LABEL THE X AND Y AXIS in your bar chart, and PROVIDE A KEY to your pie chart. Color is a great way to do this, but you can also use shading or fill.

The table on the previous page is called a 2-way table. Add totals to each row and column. Use those values in #3.

1. Fill in the blanks for each calculation on the answer sheet from the 2-way table.

In #3, a is called the **MARGINAL RELATIVE FREQUENCY**. It is the percentage of the total that chose an outcome. b is called the **JOINT RELATIVE FREQUENCY**. It is the percentage of the total that chose a combination of two outcomes. c is called the **CONDITIONAL RELATIVE FREQUENCY**. It is the percentage of students who chose one outcome who chose the second outcome.

When you have two variables, it is possible one of them can predict the other. We call these **EXPLANATORY** and **RESPONSE** variables. Explanatory variables predict, response variables are predicted. BE CAREFUL HERE. We are not saying one affects or causes the other. Correlation does not imply causation.

1. Go to the website <https://stapplet.com> and choose the “Two Categorical Variables” option. Draw a Side-By-Side Bar Graph, a Segmented Bar Graph and a Mosaic Plot in the space provided. Don’t forget to label your axes.

Note that if you stacked the side-by-side bars of the one graph, and scaled it to percentage, it would look like the segmented bar graph. If there was no association between favorite core and favorite elective, the bars of the segmented bar graph would look roughly the same.

1. Given they do not, fill in the blanks with a possible association you can see in the graph.

Notice the Mosaic plot looks similar to the Segmented Bar Graph, but Math is wider than English. This is because the Mosaic plot uses width to show relative size. 16 students out of 30 chose Math, so it is wider than English, which only 14 students chose.

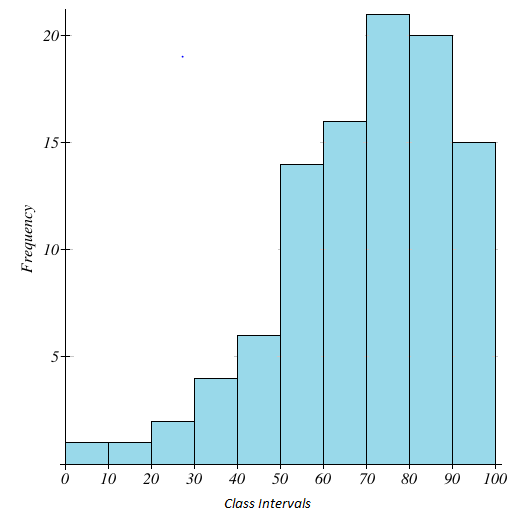
***Topic 2 – Displaying Quantitative Data***

I did another poll last fall which asked my students how many pairs of shoes they own. The list of answers they gave me is below.

2,3,3,3,3,4,4,4,5,6,6,6,7,7,8,8,9,9,9,10,10,10,11,13,17,22,34

1. Go to the website <https://stapplet.com> and choose the “One Quantitative Variable” option. Make sure “raw data” is chosen, and enter the data exactly as above. Draw and label a dotplot, stemplot, and histogram. Make sure all axes are labeled and a key is provided for the stemplot.

This data is **QUANTITATIVE**, because it has mathematical meaning and can be put in numerical order. (For example, above, it ranges from least to greatest.) Quantitative data can be **DISCRETE** (comes in chunks, like counting how many pairs of shoes you own) or **CONTINUOUS** (something you measure, which can take any value, like height). When displayed in a plot or graph, the shape and characteristics of the data has meaning. We describe quantitative data using the acronym SOCCS. Each letter stands for a characteristic you need to mention.

**S – Shape** – The shape is the location (and quantity) of the peaks of the data, whether it is symmetrical, or whether it has skew. Skew is when data trails off more sharply on one side than the other. We say the data is skewed toward its less steep side. For example, the distribution in the picture to the right is skewed left.

1. Answer the question on the answer sheet.

**O – Outliers** – Does the data have any easy to see outlying variables? What and where are they? We will learn a better way to calculate and test for this later.

1. Answer the question on the answer sheet.

**C – Center** – Where is the center of the data? This can be reported using the **median** or the **mean**. The mean is the calculated average, and the median is the middle value counting in from the sides. When there are more outliers or more skew, the median may be a better choice than the mean. Both can be found in Stapplet or your calculator.

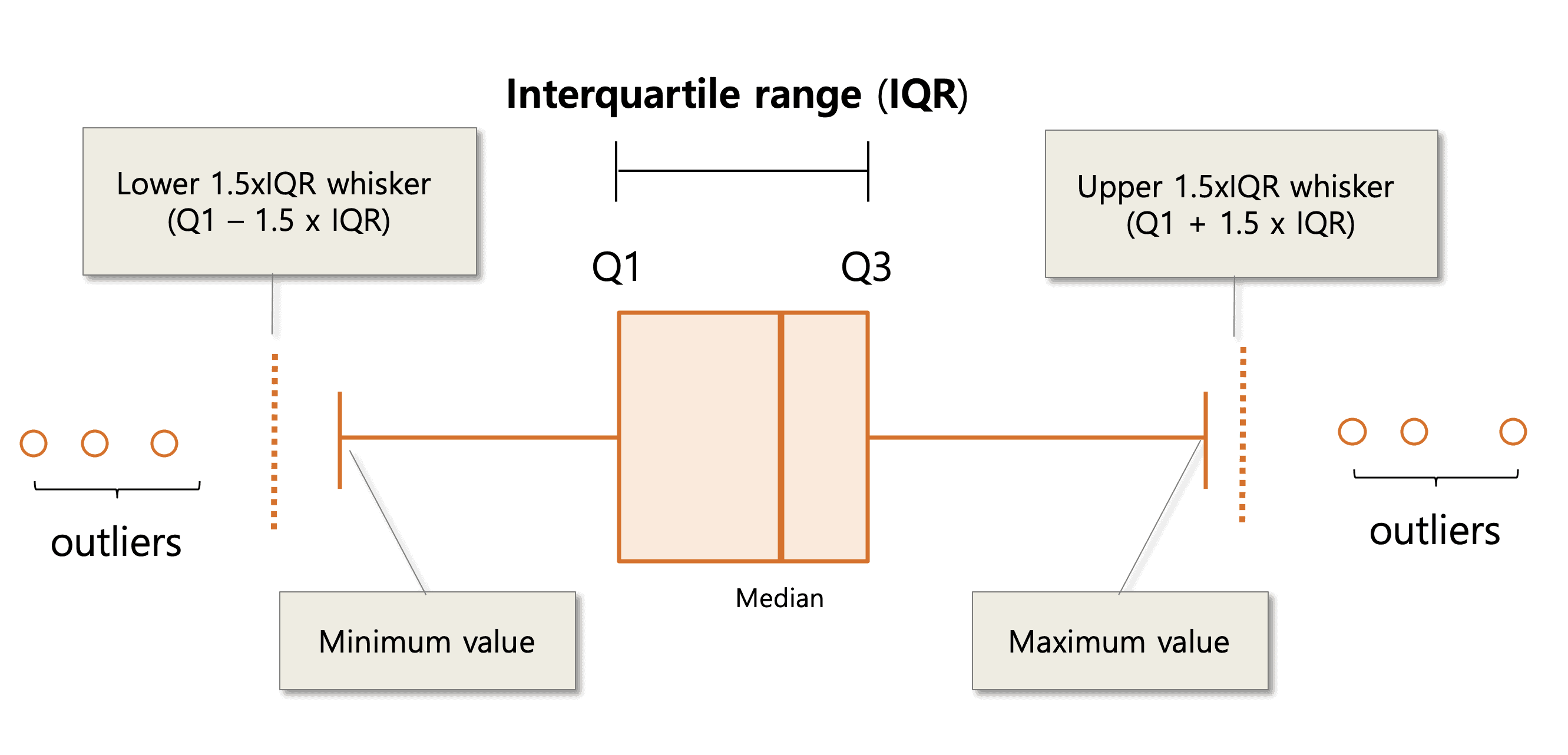
1. Answer the question on the answer sheet.

**C – Context** – This is the free one if you are paying attention. Always describe your data. Never use words like “the distribution” or “the data”. Be specific, what are you measuring? Use phrases like “The mean of the number of shoes” or “the distribution of the number of pairs of shoes”.

**S – Spread** – The easy way to describe the spread of your data is to give the range. Range is simply the difference between the highest and lowest numbers. However, range can be affected strongly by outlying data, so its not the best tool. We will learn better ways to measure spread later.

1. Answer the question on the answer sheet.

When asked to describe a distribution, you need to make sure you cover ALL the points above. Remember, **SOCCS!**

*****Topic 3 – Boxplots and the IQR***

There is another way to visualize quantitative data. It is called a boxplot. Last year, I did a quick poll of heights of students in inches. Here’s the data:

68,72,61,62,63,63,64,64,59,62,  
61,60,65,62,57,77,62,71,65,62,70

To make a boxplot, we need to find 5 key numbers, but to find them, we need to put them in order.

1. Rewrite the list of numbers in the given space in order from least to greatest.
2. As we move through the next section, follow the instructions to create your boxplot.
   1. Median – As before, the median is the “middle” number in your data. If you have an even amount of numbers, take the average of the middle two. Put a dot above the number line at the median.
   2. Max – Put a dot above the number line at the highest number in your data.
   3. Min – Put a dot above the number line at the lowest number in your data.
   4. Quartile 1 and 3 – In step a, you found and marked your Median. The quartiles are just the medians of each half of the data. Take the numbers less than the Median and count inward from both ends to find the middle number. If there is an even amount of numbers (there will be in this case because there are 10), average the middle two. This number is Q1. Put a dot above the number line here. Do the same thing for numbers greater than the Median. This number is Q3. Put a dot above the number line here.
   5. With all 5 numbers, you can make your box. Draw vertical lines at Q1, the Median, and Q3, then connect the tops and bottoms of these lines to make a box with the Median line inside it. Look at the picture above if you need an example.

Let’s pause here and look at an alternate measure of spread. In the previous section, we said one of the flaws with the Range is that it can be made much larger by extreme values. A potentially better measure is called the **Interquartile Range (IQR)**. The IQR is a measure of the distance from Q1 to Q3. It is less likely to be affected by outlying data.

Even better, IQR can help us identify outliers that might be affecting our data set. We follow a rule called 1.5\*IQR when looking for outliers. Take the IQR and multiply it by 1.5. Any number that is more than Q3 + 1.5\*IQR or less than Q1-1.5\*IQR is considered an outlier.

1. Check your data for outliers. Outliers are marked with dots on the boxplot, but not part of the boxplot itself. Mark every outlier with a dot, then (if you marked an outlier) put a new dot on the highest and lowest value that is NOT an outlier (this will be the border of your boxplot).
2. Now we draw the “whiskers” to the boxplot. From the middle of the side of the box, draw a line out to the highest non-outlier value, and the lowest non-outlier value. Draw vertical lines at these points.
3. Go to the website <https://stapplet.com> and choose the “One Quantitative Variable” option. Enter the data and choose boxplot from the dropdown menu. Yours should match the one created by Stapplet.

A boxplot lets us view the shape of the data without a distribution. What will happen if one side of the median is bunched very tightly around similar values, and the other is spread out? Well, the side that is bunched tightly will have a smaller portion of the box (and a smaller whisker). Another way to think of it is that the data is now divided into 4 main chunks. Min to Q1, Q1 to Median, Median to Q3, and Q3 to Max. Each chunk will contain 25% of the data, and its relative size on the boxplot shows how spread out that data is.

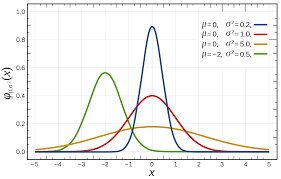
***Topic 4 – Standard Deviation***

I took a small poll of students in my class last year to see how many colleges they were applying to, and I received the following numbers.

5,4,6,5,3,6,8,3

1. Calculate the Mean, the Median, and the Range of the data.

We mentioned earlier that Range isn’t the best measure of spread. Notice that in this data, the Mean and Median are the same. This implies that the data is roughly symmetric around the center. When we have symmetric data, we have an even better measure of spread called **Standard Deviation**. Think of Standard Deviation as the “average” distance away from the mean. The more data that is close to the mean, the smaller the Standard Deviation will be.

Look at the picture to the right. It uses the Greek letter µ (Mu, pronounced Mew, like the Pokemon) to represent the Mean, and σ (Sigma) to represent the Standard Deviation. See how the spread increases as the SD gets bigger?

1. Fill in the table to calculate the Standard Deviation for the data.

Go to the website <https://stapplet.com> and choose the “One Quantitative Variable” option. Enter the numbers, and scroll down to Summary Statistics. Compare our results to theirs.

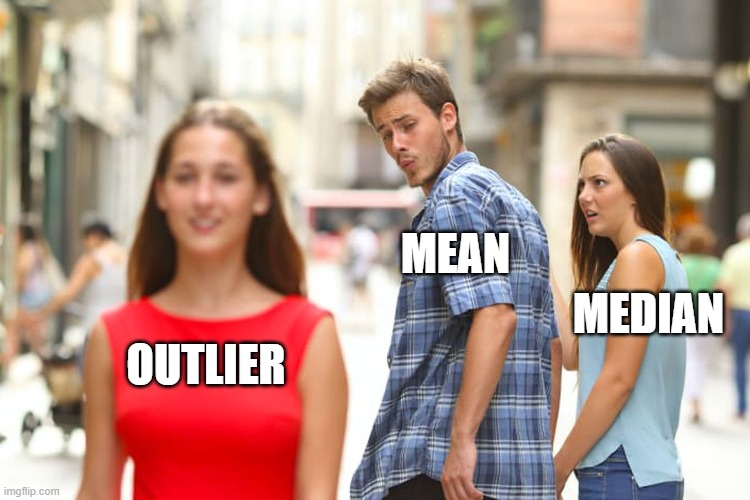
Notice that the Standard Deviation we calculated is smaller than the one that Stapplet calculated. This is because the number you calculated by hand is the standard deviation if we had data for everybody. Because we only have these 8 results, our Standard Deviation is likely too small (we would have more variation). This is a complicated calculation, but it turns out that dividing by n-1 instead of n is enough to correct for this problem. In the future, we will let the calculator or Stapplet give us the correct value.

One other note is that statisticians used to use something called **Variance.** Notice the last step of the calculation of SD was to take a square root? The number before you take the square root is the Variance. Before computer calculators, square roots were complicated and annoying. For good math reasons we will learn later, it’s easier to do math on Variances than SD’s, so to avoid having to do lots of roots, they would just leave the number in the Variance. It’s less straightforward to understand that way, and since we all have calculators, we will be primarily using SD. All you need to remember is that Variance is the square of the SD.

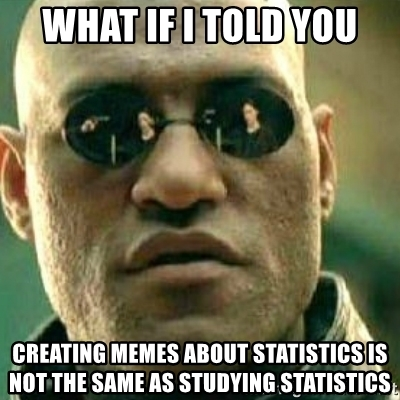
Now we have several different calculations we can report when asked to talk about distributions. Let’s summarize their good and bad points. But the short explanation is that if you have data that is symmetric and without outliers, Mean and Standard Deviation will always be the better choice. If you have skewed data, or significant outliers, Median and IQR will give you a better picture of what is happening with the distribution.

|  |  |  |
| --- | --- | --- |
| **Name** | **Good** | **Bad** |
| Median | * Simple to find manually * Resistant to skew | * Not useful in calculations |
| Mean | * Straightforward to calculate * Useful in calculations | * Not resistant to skew and outliers. |
| Range | * Very easy to calculate * Gives a useful spread of data | * Strongly affected by outliers * Not useful in calculations |
| IQR | * Not affected by outliers * Gives both spread and shape | * Annoying to calculate manually * Not mathematically useful |
| Standard Deviation | * Most useful in calculations * Accurately includes all data | * Difficult to calculate * Affected badly by skew |

Let’s finish up by seeing how Mean, Median, Standard Deviation, and IQR are affected by a strong outlier. In my data about colleges, let’s say that I polled one more student who is applying to an amazing 40 colleges.

1. Fill in the values from the previous section for Mean, Median, Standard Deviation, and IQR next to “Old”. Go back to the website <https://stapplet.com> and edit the previous data to include another data point of 40. Fill in the appropriate spots on the table for the “New” values.

Notice the large swings that the Mean and Standard Deviation take with the new value included. This is what it means to be “non-resistant” to outlier data.

You’ve reached the end of the AP Stat Summer Assignment. Great job! The purpose of this was to prepare you to enter the class ready to go to our first big statistical concept, Z-scores. You will be turning in the answer page of this work to me for a test grade. Statistics is a little different from some math classes in that it builds consecutively, layer on layer. The first week of class you will learn to calculate a z-score, and I guarantee you will need to calculate a z-score on the last week of class as well. The difference will be the amount of meaning that calculation carries.

When we start class, I will be assuming you have worked on, read through, and generally understood the concepts on this assignment. They will show up throughout the year, so please make sure you ask questions if something is unclear. My email is [jhoiler@ccboe.com](mailto:jhoiler@ccboe.com) and you can email me during the summer. I will do my best to respond within a day or two. I look forward to working with you in the fall.

***For more help, here’s some links to helpful resources.***

Kahn Academy’s Get Ready for AP Stat Page - [Get ready for AP® Statistics | Math | Khan Academy](https://www.khanacademy.org/math/get-ready-for-ap-statistics)

Statistics 1 Playlist - <https://www.youtube.com/watch?v=MXaJ7sa7q-8&list=PL0KQuRyPJoe6KjlUM6iNYgt8d0DwI-IGR>

(I recommend videos 1-8 in the playlist. You will actually be ahead at that point.)

Statistics Crash Course Playlist - <https://www.youtube.com/watch?v=zouPoc49xbk&list=PL8dPuuaLjXtNM_Y-bUAhblSAdWRnmBUcr>

Name \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ **AP Stat Summer Assignment Answer Sheet**

1. Elective Table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Art | Music | P.E. | Language | Technology |
| Total |  |  |  |  |  |

1. Draw Here
2. Fill in the blanks below
   1. \_\_\_\_\_\_\_\_\_\_ / \_\_\_\_\_\_\_\_\_ = \_\_\_\_\_\_\_\_\_\_\_\_

(chose PE) (total students)

* 1. \_\_\_\_\_\_\_\_\_\_ / \_\_\_\_\_\_\_\_\_ = \_\_\_\_\_\_\_\_\_\_\_\_

(chose Math and Art) (total students)

* 1. \_\_\_\_\_\_\_\_\_\_ / \_\_\_\_\_\_\_\_\_ = \_\_\_\_\_\_\_\_\_\_\_\_

(chose Math and Tech) (chose Math)

1. Draw Here

1. If you choose \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ you are more likely to choose \_\_\_\_\_\_\_\_\_\_\_\_\_.
2. Draw Here
3. For our shoe data, the dot plot and histogram are skewed to the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.
4. For our shoe data, there is a potential outlier at \_\_\_\_\_\_\_\_\_\_\_\_.
5. The mean of the data is \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_. The median of the data is \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.
6. The range of the data is \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.
7. Rewrite the numbers here.
8. (through 15) Follow the instructions to create the boxplot.



1. Fill in the boxes

|  |  |  |
| --- | --- | --- |
| **Mean** | **Median** | **Range** |
|  |  |  |

1. Follow the directions to fill in the table. I did the first line for you.

|  |  |  |
| --- | --- | --- |
| **Value** | **Value - Mean** |  |
| 5 | 5-5=0 |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  | (Add column 3) TOTAL 🡪 |  |
|  | (Divide by 8) AVERAGE 🡪 |  |
|  | (Take the Square Root) SD 🡪 |  |

1. Fill in the table with the results from Stapplet.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Median** | **Mean** | **IQR** | **Standard Deviation** |
| **Old** |  |  |  |  |
| **New** |  |  |  |  |